BioSystems 101 (2010) 127–135

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

The role of internal duplication in the evolution of multi-domain proteins

J.C. Nacher[a,*,1], M. Hayashida[b,1], T. Akutsu[b,**]

[a] Department of Complex and Intelligent Systems, Future University-Hakodate, 116-Kamedankano, Hakodate, Hokkaido 041-8655, Japan
[b] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Japan

ARTICLE INFO

Article history:
Received 17 March 2010
Received in revised form 24 May 2010
Accepted 25 May 2010

Keywords:
Protein domain
Genome evolution
Evolutionary dynamics
Systems obeying scaling-laws
Mathematical model

ABSTRACT

Many proteins consist of several structural domains. These multi-domain proteins have likely been generated by selective genome growth dynamics during evolution to perform new functions as well as to create structures that fold on a biologically feasible time scale. Domain units frequently evolved through a variety of genetic shuffling mechanisms. Here we examine the protein domain statistics of more than 1000 organisms including eukaryotic, archaeal and bacterial species. The analysis extends earlier findings on asymmetric statistical laws for proteome to a wider variety of species. While proteins are composed of a wide range of domains, displaying a power-law decay, the computation of domain families for each protein reveals an exponential distribution, characterizing a protein universe composed of a thin number of unique families. Structural studies in proteomics have shown that domain repeats, or internal duplicated domains, represent a small but significant fraction of genome. In spite of its importance, this observation has been largely overlooked until recently. We model the evolutionary dynamics of proteome and demonstrate that these distinct distributions are in fact rooted in an internal duplication mechanism. This process generates the contemporary protein structural domain universe, determines its reduced thickness, and tames its growth. These findings have important implications, ranging from protein interaction network modeling to evolutionary studies based on fundamental mechanisms governing genome expansion.

1. Introduction

In proteins, modularity emerges at several levels overlapping the standard categories of tertiary and quaternary protein structure (Wetlaufer, 1973; Doolittle, 1995). In the most established view, regions of consecutive residues in polypeptide chains that fold into distinct structural modules are called *protein domains* and are responsible not only for the cohesion between side-chains but also for multiple biological functions related to protein–protein and cell–cell interactions, signal transduction and cell differentiation processes (Janin and Chothia, 1985). It is largely recognized that evolutionarily conserved protein domains may occur alone, although many proteins contain two or more domains and the largest multi-domain proteins consist up to a few hundreds of domains (Wetlaufer, 1973; Li et al., 2001). Yet another perspective introduces a definition of protein modules considered as more compact structural units in proteins with a length in the range of 20–40 residues (Go, 1983, 1981). However, even though currently there is a large variety of protein structure information collected and stored in databases, we do not yet have a simple answer to the principal question formulated in Doolittle's seminal work in 1995 on multiplicity of domains (Wetlaufer, 1973; Doolittle, 1981): *how many domain families exist and what were their origin?*

While former ideas were based on the assumption of a large number of domains devised early in evolutionary history that generated the current protein universe by means of domain-shuffling events (Dorit et al., 1990), another perspective is that in the early stages of life, a more reduced number of small polypeptides acted as seeds for the most modern multi-domain proteins that evolved from them by means of a variety of genetic processes (Doolittle, 1981) like fusion and fission of domains (Kummerfeld and Teichmann, 2005), and mutation and duplication of proteins (Wetlaufer, 1973; Doolittle, 1981). Although the dynamics of genome growth through shuffling of protein domains have been studied extensively over decades, recent experimental observations of a significantly large number of *domain repeats* of several domains from the same family, mainly in numerous eukaryotic proteins, suggest that one process involving domain recombination may have been overlooked (Moore, 2008; Björklund et al., 2006). While the exact mechanism behind the existence of domain repeats is not yet fully understood, it has been argued that domain repeats could have been generated by a so-called partial *internal duplication* process within a gene, where a region or protein domain is

* Corresponding author. Tel.: +81 0138 34 6123.
** Corresponding author. Tel.: +81 0774 38 3015.
E-mail addresses: nacher@fun.ac.jp (J.C. Nacher), takutsu@kuicr.kyoto-u.ac.jp (T. Akutsu).
[1] These authors equally contributed to this work.

0303-2647/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2010.05.005

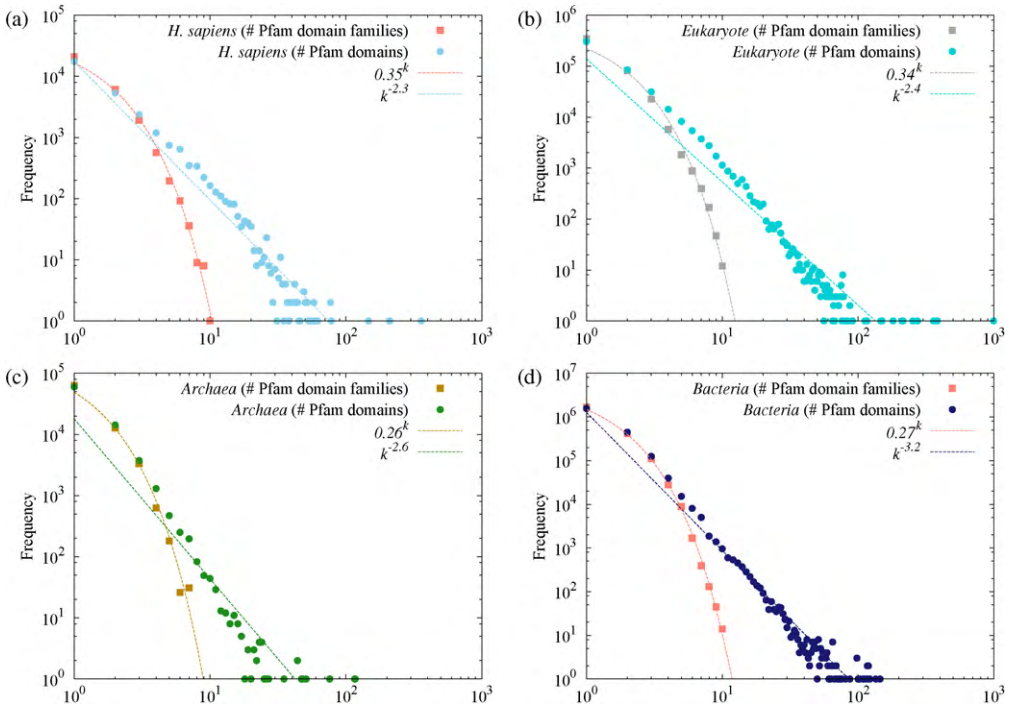

Fig. 1. Distributions of the number of domain families (squares) and the number of domains (dots) in a protein. (a) *H. sapiens*, (b) 68 eukaryotic organisms, (c) 56 archaeal organisms and (d) 929 bacterial organisms.

duplicated and allocated next to its origin. When all the unique domains (we refer to them hereafter as domain families, Moore, 2008) are copied, the process is called a complete internal duplication event (Zhang, 2003). This internal duplication process that occurs within a protein is also known as *tandem duplication* within a gene and should not be confused with the usual external protein duplication process, where two individual proteins share the same genetic material after being duplicated. It is believed that tandem duplication events might explain the frequent observation of several domain repeats from the same family in eukaryotic genomes (Moore, 2008; Björklund et al., 2006).

Here we examine the protein domain statistics retrieved from Pfam, SMART, Gene3D, ProDom and TIGRFAMs databases and consisting of 68 eukaryotic, 56 archaeal, and 929 bacterial organisms. We show that this analysis confirms earlier observations (Koonin et al., 2002; Beyer and Wilhelm, 2005) and extends them to numerous organisms in the three kingdoms of life. The results show that the number of total protein domains and the number of domain families in a protein are governed by different statistical laws. While the former follows a power-law distribution, the latter exhibits an exponential statistics. We develop a methodology and propose an evolutionary dynamics model, based on a rate equation formalism, and consisting of domain fusion, mutation, protein duplication and internal duplication processes. We then demonstrate that these distinct distributions are in fact rooted in the internal duplication mechanism. The analytical results derived from the evolutionary dynamics model as well as computer simulation show that this *domain-repeats event* generates a wide number of domains in a protein while at the same time preserving a thin number of domain families across proteome species. To our knowledge, this is the first mathematical model of protein domain evolution that explicitly takes into account the effect of internal duplication mechanism and provides analytical solution. Taken together, our findings offer insight into the fundamental mechanisms governing genome expansion with potential implications ranging from protein interaction network modeling to evolutionary studies based on fundamental mechanisms governing the genome expansion.

2. Statistical analysis

2.1. Proteome datasets

The proteome data was collected in UniProt format from the Integr8 database (version 90) (Kersey et al., 2005). UniProt format (The Uniprot Consortium, 2008) includes pointers to the entries of the Pfam database (Finn, 2008) related to each protein, and the number of copies of the same Pfam domain in a protein. This information was used to obtain the number of domain families as well as the number of domains in a protein. The Integr8 database (version 90) includes 68 eukaryotic organisms, 57 archaeal organisms, and 932 bacterial organisms. We excluded one archaeal organism and three bacterial organisms that contained proteins from viruses in their proteomes, and obtained the distribution of the number of domain families and the number of domains in a protein for all eukaryotic, archaeal, and bacterial organisms (see Fig. 1). For each of 18 organisms, Table S1 in Supplementary Information (SI) shows the total number of proteins, the total number of proteins having Pfam domains, the total number of domain families and the total number of domains. It also shows the best-fit results on the distribution of the number of domain families in a protein by an exponential function and that of the number of domains in a protein by a power-law function. A complementary analysis was also performed using SMART, Gene3D, ProDom and TIGRFAMs databases (see Figs. S2–S5).

2.2. The contemporary protein domain universe is wide but thin

In this work, we have investigated the impact of the tandem duplication process in protein evolution and found that the size of the modern protein domain universe has been shaped and drastically constrained by an often ignored genetic mechanism.

The protein domain statistics were investigated and the results confirmed and extended earlier observations (Koonin et al., 2002) to a wider variety of organisms. The data analysis showed that while the probability to find a number of domains k in a protein follows a